Session 2: Reading data from text files

In this session we will enter data into SAS from a text file using the `infile` statement.

Consider the leukemia data set `leukemia.dat`. This is a text file containing three variables: `trt`, `remiss` and `status`, containing information on treatment arm, remission time and status (1=event, 0=censored observation).

Before we present the data step consider how format information is generated in SAS. It is always a good idea not to use character fields in data, but rather use numerical codes and then associate these codes with a format (like the `label` command in STATA).

Formatting is done through `PROC FORMAT`. In the following code we create two formats: One for the censor indicator called `status` (it does not matter that the name of the format is the same as the name of the variable `status`) and one for the treatment group called `trtfmt` as follows:

```
proc format;
     value trtfmt 0='Control' 1='6-MP';
     value status 0='Censored' 1='Event';
run;
```

The following components of the format procedure are important. First the heading

```
proc format;
```

The heading takes no options. Then follows the word `value`, which is followed by the name of the format, which is followed in turn by the numerical fields that are equated with their character interpretations in (single or double) quotes. Note that there is no semicolon until the definition of the format is complete.

Here we have created two formats

```
     value trtfmt 0='Control' 1='6-MP';
     value status 0='Censored' 1='Event';
```

Thus, trtfmt assigns the word “control” to 0 and “6-MP” to 1, while the format status assigns the word “censored” to 0 and “event” to 1. This is an example of a numerical format. You can easily have character formats. For example, suppose that you have data where “Y” is yes and “N” is no. Then a character format assigning “N” to “No” and “Y” to “Yes” is as follows:

```
     value $yesno 'Y'='Yes' 'N'='No';
```

Notice the “$” before the value `yesno`, denoting that this is a character format.

Running the format procedure we get the following comments in the log file:

```
1    proc format;
2         value trtfmt 0='Control' 1='6-MP';
NOTE: Format TRTFMT has been output.
3         value status 0='Censored' 1='Event';
NOTE: Format STATUS has been output.
4    run;

NOTE: PROCEDURE FORMAT used:
      real time           0.15 seconds
      cpu time            0.03 seconds
```

Be aware that, at this point, no format has been attached to any variable in any data set. Now let's enter the leukemia data set. The code is as follows:

```
data leukemia;
     infile 'leukemia.dat';
     input trt remiss status;
     format trt trtfmt. status status.;
     label remiss='Time to end of remission'
           status='Censoring indicator'
           trt='Treatment assignment';
run;
```

Concentrate on the following statement:

```
     infile 'leukemia.dat';
```

This indicates to SAS which file to get the data from. There should be some path information available inside the quotes. For example, if the file leukemia.dat (which *must* be a text file) is in a diskette, you should write

```
     infile 'a:\leukemia.dat';
```

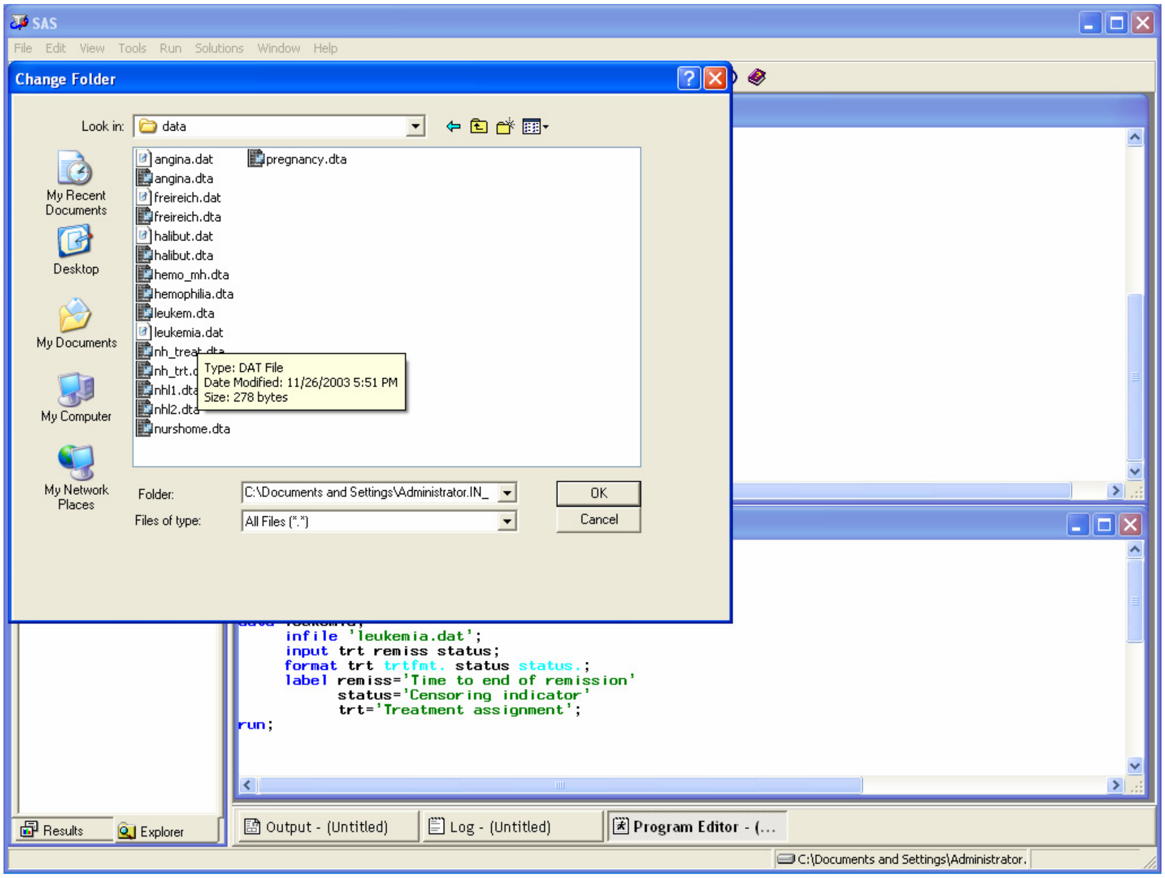
Alternatively you can make the directory where the file is located the *default* directory for SAS. The default directory is located at the bottom right corner of the SAS window.

You can change it by double-clicking the mouse on it and using the Windows change folder window to search for the appropriate directory archive (see next page). Once you find the appropriate directory you choose it by pressing the OK button.

Then you can run the data step being certain that SAS will look in the appropriate directory for the data file. Now to assign the appropriate formats to the appropriate variables you should write

```
     format trt trtfmt. status status.;
```

Notice that, to indicate which word represents a format, you follow the format name with a period ("."). If more than one variable have the same format, you can have a list of variables followed by their common format name.



Look in: data

Type: DAT File
Date Modified: 11/26/2003 5:51 PM
Size: 278 bytes

Folder: C:\Documents and Settings\Administrator.IN_

Files of type: All Files (*.*)

```
    infile 'leukemia.dat';
    input trt remiss status;
    format trt trtfmt. status status.;
    label remiss='Time to end of remission'
          status='Censoring indicator'
          trt='Treatment assignment';
run;
```

The second novel part of the data step is the `label` statement.

```
      label remiss='Time to end of remission'
            status='Censoring indicator'
            trt='Treatment assignment';
```

This assigns a name to the variables, which will appear in graphs and tables and will provide more information for the variable than its SAS name (and will make your output friendlier to non-statisticians).

Running the data step we have the following output in the log file:

```
6    data leukemia;
7         infile 'leukemia.dat';
8         input trt remiss status;
9         format trt trtfmt. status status.;
10        label remiss='Time to end of remission'
11              status='Censoring indicator'
12              trt='Treatment assignment';
13   run;

NOTE: The infile 'leukemia.dat' is:

      File Name=C:\Documents and Settings\cyiannou\Desktop\BIO223 (Survival-
      Yiannoutsos)\data\leukemia.dat, RECFM=V,LRECL=256

NOTE: 42 records were read from the infile 'leukemia.dat'.
      The minimum record length was 5.
      The maximum record length was 7.
NOTE: The data set WORK.LEUKEMIA has 42 observations and 3 variables.
NOTE: DATA statement used:
      real time           0.74 seconds
      cpu time            0.10 seconds
```

Printing the data set we get

```
proc print data=leukemia label;
     title 'Leukemia data set';
run;
```

Notice the option `label` in the invocation of the procedure. This will make the variables to be headed by their label rather than their SAS name. The output is as follows:

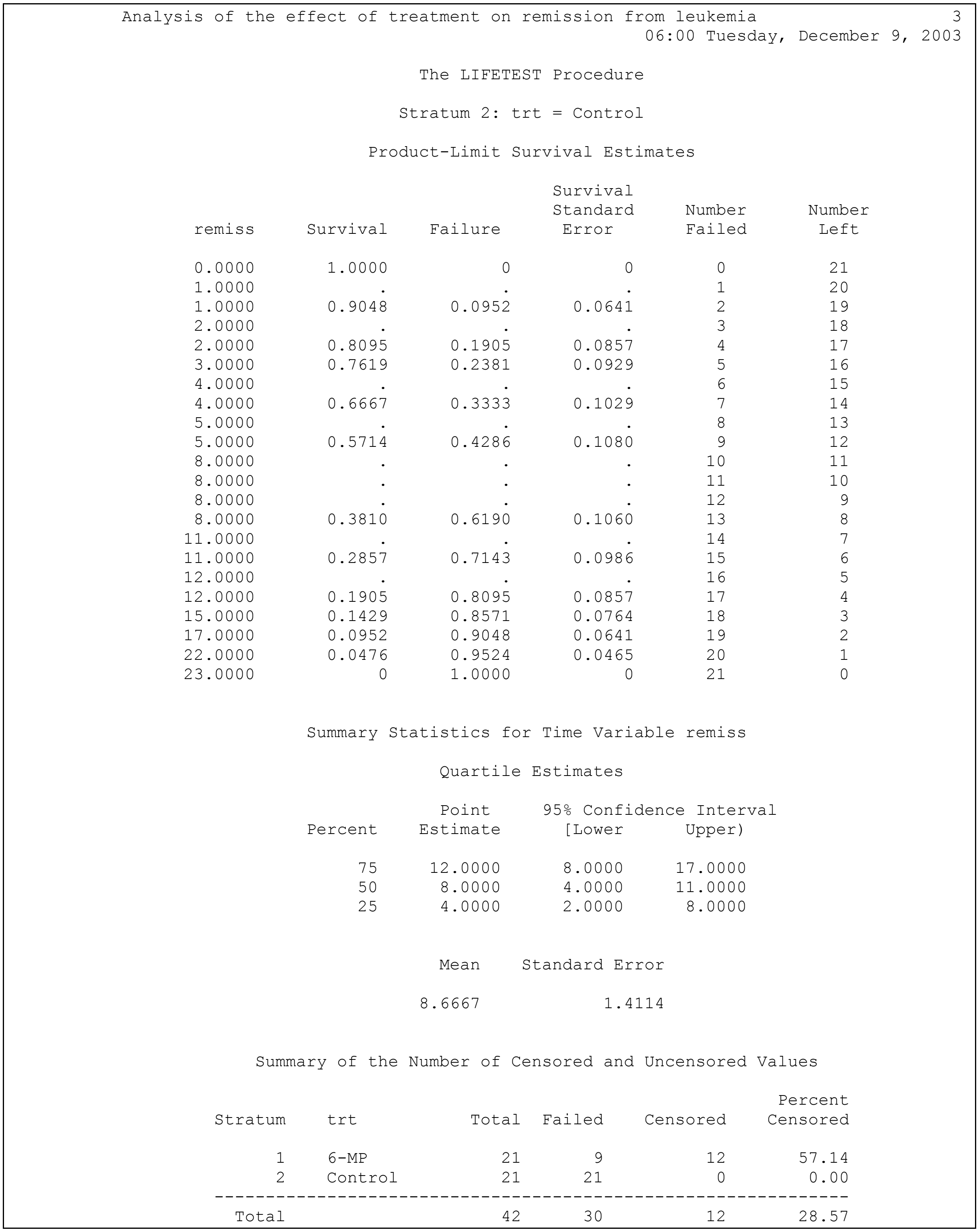
```
                                         Time to
                    Treatment             end of      Censoring
             Obs    assignment          remission     indicator

               1     Control                1         Event
               2     Control                1         Event
               3     Control                2         Event
               4     Control                2         Event
               5     Control                3         Event
               6     Control                4         Event
               7     Control                4         Event
               8     Control                5         Event
               9     Control                5         Event
              10     Control                8         Event
              11     Control                8         Event
              12     Control                8         Event
              13     Control                8         Event
              14     Control               11         Event
              15     Control               11         Event
              16     Control               12         Event
              17     Control               12         Event
              18     Control               15         Event
              19     Control               17         Event
              20     Control               22         Event
              21     Control               23         Event
              22     6-MP                   6         Censored
              23     6-MP                   6         Event
              24     6-MP                   6         Event
              25     6-MP                   6         Event
              26     6-MP                   7         Event
              27     6-MP                   9         Censored
              28     6-MP                  10         Censored
              29     6-MP                  10         Event
              30     6-MP                  11         Censored
              31     6-MP                  13         Event
              32     6-MP                  16         Event
              33     6-MP                  17         Censored
              34     6-MP                  19         Censored
              35     6-MP                  20         Censored
              36     6-MP                  22         Event
              37     6-MP                  23         Event
              38     6-MP                  25         Censored
              39     6-MP                  32         Censored
              40     6-MP                  32         Censored
              41     6-MP                  34         Censored
              42     6-MP                  35         Censored
```

Now let's run the `lifetest` procedure in order to carry out the Kaplan-Meier analysis on this data set, *stratifying* it by treatment assignment and carrying out the log-rank test. The SAS code is as follows:

```
proc lifetest data=leukemia method=pl plot=(s);
     time remiss*status(0);
     strata trt;
     title 'Analysis of the effect of treatment on remission from leukemia';
run;
```

The output is as follows:

```
        Analysis of the effect of treatment on remission from leukemia                  2
                                                            06:00 Tuesday, December 9, 2003

                                        The LIFETEST Procedure

                                        Stratum 1: trt = 6-MP

                                   Product-Limit Survival Estimates

                                                      Survival
                                                      Standard     Number      Number
                  remiss     Survival    Failure       Error       Failed       Left

                  0.0000       1.0000            0           0        0          21
                  6.0000            .            .           .        1          20
                  6.0000            .            .           .        2          19
                  6.0000       0.8571       0.1429      0.0764        3          18
                  6.0000*           .            .           .        3          17
                  7.0000       0.8067       0.1933      0.0869        4          16
                  9.0000*           .            .           .        4          15
                 10.0000       0.7529       0.2471      0.0963        5          14
                 10.0000*           .            .           .        5          13
                 11.0000*           .            .           .        5          12
                 13.0000       0.6902       0.3098      0.1068        6          11
                 16.0000       0.6275       0.3725      0.1141        7          10
                 17.0000*           .            .           .        7           9
                 19.0000*           .            .           .        7           8
                 20.0000*           .            .           .        7           7
                 22.0000       0.5378       0.4622      0.1282        8           6
                 23.0000       0.4482       0.5518      0.1346        9           5
                 25.0000*           .            .           .        9           4
                 32.0000*           .            .           .        9           3
                 32.0000*           .            .           .        9           2
                 34.0000*           .            .           .        9           1
                 35.0000*           .            .           .        9           0

                      NOTE: The marked survival times are censored observations.


                             Summary Statistics for Time Variable remiss

                                          Quartile Estimates

                                          Point     95% Confidence Interval
                             Percent    Estimate      [Lower      Upper)

                                  75        .         23.0000       .
                                  50     23.0000      13.0000       .
                                  25     13.0000       6.0000     23.0000


                                          Mean    Standard Error

                                       17.9092           1.6474

NOTE: The mean survival time and its standard error were underestimated because the largest
observation was censored and the estimation was restricted to the largest event time.
```



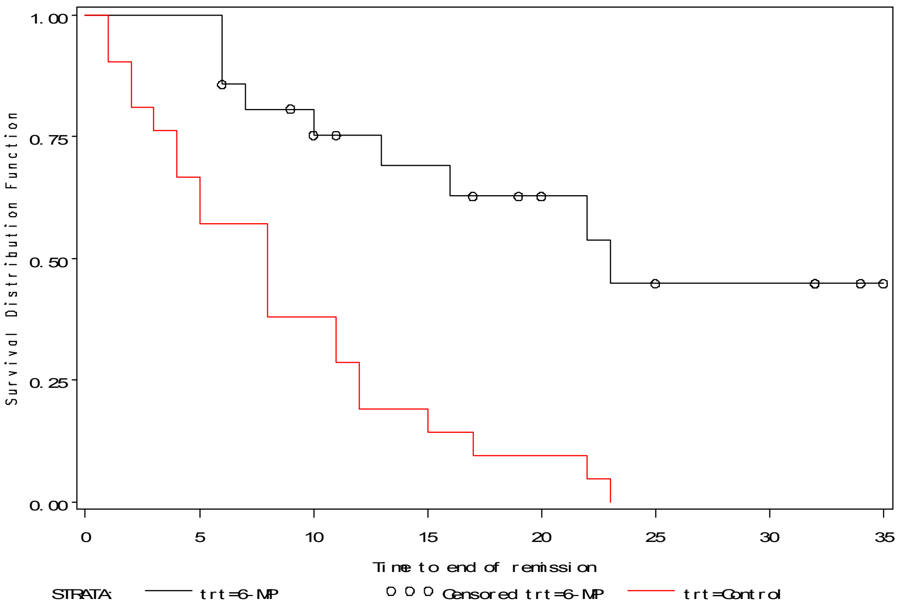
```
           Analysis of the effect of treatment on remission from leukemia            3
                                                    06:00 Tuesday, December 9, 2003

                                The LIFETEST Procedure

                               Stratum 2: trt = Control

                            Product-Limit Survival Estimates

                                                 Survival
                                                 Standard     Number      Number
                  remiss     Survival    Failure     Error       Failed       Left

                  0.0000       1.0000           0          0          0          21
                  1.0000            .           .          .          1          20
                  1.0000       0.9048      0.0952     0.0641          2          19
                  2.0000            .           .          .          3          18
                  2.0000       0.8095      0.1905     0.0857          4          17
                  3.0000       0.7619      0.2381     0.0929          5          16
                  4.0000            .           .          .          6          15
                  4.0000       0.6667      0.3333     0.1029          7          14
                  5.0000            .           .          .          8          13
                  5.0000       0.5714      0.4286     0.1080          9          12
                  8.0000            .           .          .         10          11
                  8.0000            .           .          .         11          10
                  8.0000            .           .          .         12           9
                  8.0000       0.3810      0.6190     0.1060         13           8
                 11.0000            .           .          .         14           7
                 11.0000       0.2857      0.7143     0.0986         15           6
                 12.0000            .           .          .         16           5
                 12.0000       0.1905      0.8095     0.0857         17           4
                 15.0000       0.1429      0.8571     0.0764         18           3
                 17.0000       0.0952      0.9048     0.0641         19           2
                 22.0000       0.0476      0.9524     0.0465         20           1
                 23.0000            0      1.0000          0         21           0


                        Summary Statistics for Time Variable remiss

                                  Quartile Estimates

                                  Point     95% Confidence Interval
                        Percent    Estimate      [Lower      Upper)

                             75     12.0000      8.0000     17.0000
                             50      8.0000      4.0000     11.0000
                             25      4.0000      2.0000      8.0000


                                  Mean    Standard Error

                                8.6667           1.4114


                     Summary of the Number of Censored and Uncensored Values

                                                                      Percent
                    Stratum    trt           Total  Failed    Censored    Censored

                          1    6-MP             21       9          12       57.14
                          2    Control          21      21           0        0.00
                    ---------------------------------------------------------------
                      Total                     42      30          12       28.57
```

```
          Analysis of the effect of treatment on remission from leukemia                4
                                                       06:00 Tuesday, December 9, 2003

                              The LIFETEST Procedure

             Testing Homogeneity of Survival Curves for remiss over Strata


                                 Rank Statistics

                         trt         Log-Rank    Wilcoxon

                         6-MP         -10.251     -271.00
                         Control       10.251      271.00


                 Covariance Matrix for the Log-Rank Statistics

                        trt              6-MP       Control

                        6-MP          6.25696      -6.25696
                        Control      -6.25696       6.25696


                 Covariance Matrix for the Wilcoxon Statistics

                        trt              6-MP       Control

                        6-MP          5457.11      -5457.11
                        Control      -5457.11       5457.11


                           Test of Equality over Strata

                                                      Pr >
                  Test      Chi-Square      DF    Chi-Square

                  Log-Rank     16.7929       1      <.0001
                  Wilcoxon     13.4579       1      0.0002
                  -2Log(LR)    16.4852       1      <.0001
```

The log-rank test given at the bottom of the output has test statistic 16.7929, which, compared to a chi-square distribution with one degree of freedom, is statistically significant. We conclude that there is a significant difference in survival between the two treatment groups (although we cannot tell directly from the log-rank test which treatment has the advantage.

This can be gleaned from studying the medians in the two groups (8 weeks in the control group versus 23 weeks in the treatment group), or by inspection of the Kaplan-Meier plot

Analysis of the effect of treatment on remission from leukemia


Survival Distribution Function
Time to end of remission
STRATA:
trt=6-MP
○ ○ ○ Censored trt=6-MP
trt=Control